SPUR-XMC: Context-Adaptive Spurious Correlation Mitigation for Extreme Multi-Label Classification of Biomedical Literature

Ratri Mukherjee
University of Iowa
Iowa City, Iowa, USA
ratri-mukherjee@uiowa.edu

Shailesh Dahal
University of Iowa
Iowa City, Iowa, USA
shailesh-dahal@uiowa.edu

Kishlay Jha
University of Iowa
Iowa City, Iowa, USA
kishlay-jha@uiowa.edu

Abstract

Extreme multi-label text classification (XMTC) of biomedical literature aims to classify articles into a set of pre-defined labels from a large and highly skewed label distribution space. While recent pretrained language models (PLMs) based XMTC approaches have made great advances, they still struggle to generalize to tail labels that constitute concepts of high biomedical significance (e.g., rare diseases). A key factor contributing to poor tail label performance is the presence of *spurious correlations*, where model inadvertently learns non-informative feature-label associations. To address this, we propose a novel context-adaptive spurious correlation mitigation technique (namely **SPUR-XMC**) that selectively attenuates spurious influences according to the context of input article. A unique advantage of the proposed approach is that it does not modify the architectures of existing PLMs for XMTC tasks and works as an add-on enhancer module. Extensive experiments across benchmark biomedical literature datasets validate the effectiveness of proposed approach.

1 Introduction

Extreme multi-label text classification (XMTC) of biomedical literature refers to the task of assigning a biomedical article with its most relevant subset of labels from an extremely large label space (Liu et al., 2017). PubMed (Lu, 2011) is an illustrative example of XMTC, where research articles are indexed with approximately 8–12 Medical Subject Headings (MeSH) (Lipscomb, 2000) from the collection of roughly 30,000 MeSH terms. As a fundamental task in natural language processing, XMTC plays a crucial role in empowering a wide range of biomedical applications, including information retrieval, question answering, and hypothesis generation (Zhao et al., 2021). A key challenge in XMTC is to develop text classifiers that remain robust to the long-tail distribution of highly skewed label spaces. While tail labels occur in only a small fraction of biomedical articles, they account for the majority of labels in the biomedical label space (Lu, 2011; Mukherjee and Jha, 2024; Mukherjee et al., 2025). Moreover, tail labels in the biomedical literature often correspond to concepts of biomedical significance, such as rare diseases, genes, or proteins. Thus, it is imperative to develop novel XMTC approaches that achieve strong generalization performance on tail labels.

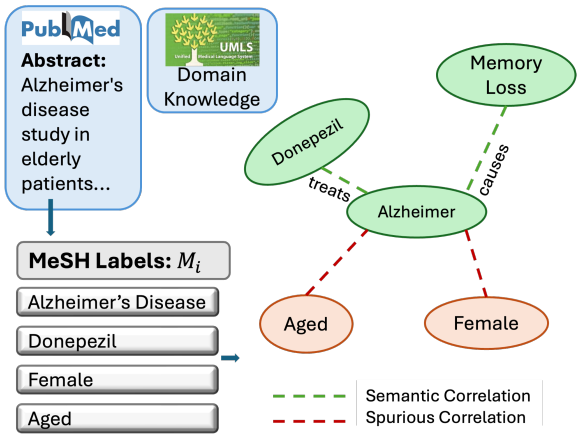


Figure 1: Abstract-level knowledge graphs identify spurious codes (orange) versus semantically connected medical concepts (green) based on UMLS relationships.

Recently, domain-specific pre-trained language models (PLMs) such as PubMedBERT (Gu et al., 2021) and BERTMeSH (You et al., 2021) have shown that they can generate high-quality feature representations of biomedical text and achieve state-of-the-art XMTC performance. Despite significant advances, existing PLMs show limited generalizability to tail or rare labels. This result can be partly attributed to the presence of *spurious correlations* in the training data (i.e., biomedical literature) used for the training of domain-specific PLMs. Specifically, in this paper, spurious correlations are referred to as the inadvertent learning of

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 34025–34045
October 24-29, 2026 ©2026 Association for Computational Linguistics

non-informative feature-label associations that can constrain generalization capability and classification performance. For example, consider Figure 1 that shows an article on the topic of "Alzheimer's disease" annotated with its corresponding concepts. While labels such as "Memory loss" and "Donepezil" exhibit meaningful semantic relationships with "Alzheimer's disease", other labels, such as "female" do not reflect meaningful associations and thus are deemed non-informative or spurious correlations. However, it should be noted that the concept "female" is not spurious globally. For instance, it is possible that "female" is semantically informative within the context of another article. Together, these examples illustrate that spuriousness is contextual and not a global characteristic. This motivates the need for an instance-level mitigation strategy as proposed in this work.

To address the above challenges, we propose a novel approach (named **SPUR-XMC**) that accurately identifies and mitigates spurious correlations in the biomedical literature. Specifically, we propose to leverage biomedical knowledge bases such as the Unified Medical Language System (UMLS) (Bodenreider, 2004) to accurately identify spurious correlations. Since the UMLS is curated and maintained by domain experts, it provides a reliable source of information for distinguishing between semantically meaningful and spurious correlations. Having identified spurious correlations, we alleviate *spuriousness* by applying a context-specific translation that remains invariant to label-frequency imbalance. Unlike existing spurious methods (Zheng et al., 2024; Han et al., 2024; Xuan et al., 2024) that mainly apply global mitigation strategies, our approach performs context-adaptive mitigation, which selectively attenuates spurious influences according to the input article and its predicted label relevance. Finally, a unique advantage of SPUR-XMC is that it does not modify the backbone architectures of PLMs used for XMTC; instead, it acts as an add-on enhancer module to boost their predictive performance. Our main contributions are summarized below:

- We address the challenge of long-tail distribution in XMTC of biomedical literature from a new perspective, i.e., mitigating spurious correlation in a context-specific manner that greatly improves the generalization performance for tail or rare labels.
- We propose a context-adaptive spurious correlation mitigation technique that adaptively modulates the spurious influences according to the context of input article. This is different from existing spurious correlation techniques that usually adopt a global mitigation strategy and are not robust to the long-tail distributions commonly observed in XMTC settings.
- Experiments conducted on benchmark biomedical literature datasets show that the proposed approach outperforms baseline algorithms across multiple evaluation metrics. We provide theoretical support for SPUR-XMC and conduct ablation studies to evaluate the contribution of model design choices.

2 Related Works

XMTC in Biomedical NLP Recent advances in biomedical XMTC include AttentionXML (You et al., 2019b), HAXMLNet (You et al., 2019a), and X-Transformer (Chang et al., 2020) that mainly leveraged transformer-based architectures. More recent biomedical XMTC approaches primarily rely on pretrained language models such as SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2020), and PubMedBERT (Gu et al., 2021), with extensions including BioELECTRA (Kanakarajan et al., 2021), BERTMeSH (You et al., 2021), MedicalBERT (Reddy et al., 2025), ClinicalModernBERT (Lee et al., 2025), BioMedLM (Bolton et al., 2024), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021). Despite their success, these models inherit corpus biases, causing spurious correlations to dominate their representations and hindering rare-label prediction.

Mitigation of Spurious Correlations Mitigating reliance on spurious features is critical for robust models. Geometric approaches such as INLP (**?**), and LEACE (Belrose et al., 2023) remove concept-specific information from representations via projection or linear transformations. Other methods identify and suppress spurious signals at the token level (Wang et al., 2022), while DFR (Kirichenko et al., 2023) shows that such correlations often reside in the classification head, suggesting robustness can be improved through appropriate reweighting. Building on these insights, we propose an add-on enhancer for BERT-based models that directly mitigates spurious correlations within the learned representation space.

3 Methodology

3.1 Preliminaries

Identifying Spurious Correlations. To accurately distinguish semantically meaningful associations from spurious ones, we choose biomedical knowledge bases such as the Unified Medical Language System (UMLS) (Lindberg et al., 1993), which provide expert-curated semantic relations among biomedical entities. For a document x_i with concept codes $M_i = \{m_1, \ldots, m_{|M_i|}\}$, we construct a document-specific subgraph $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i, \mathcal{R})$, where $\mathcal{V}_i = M_i$ and $\mathcal{E}_i$ includes knowledge graph edges linking concept pairs in M_i. Relation types $\mathcal{R}$ represent clinically meaningful relations (e.g., *treats*, *causes*, *associated with*). Concepts are then partitioned into connected and spurious nodes. Connected concepts share at least one semantic relation within the document, whereas isolated concepts lack such connections and are treated as potential spurious co-occurrences.

BERT-based Pre-trained Encoders. As a running example, our approach uses transformer-based pretrained LMs derived from the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). BERT learns contextualized token representations through stacked self-attention layers. Given an input document $x = \{w_1, \ldots, w_n\}$, the tokenizer converts the text into subword tokens, which are embedded using token, positional, and segment embeddings. The transformer encoder then produces contextual representations: $H = \text{BERT}(x) = \{h_{\texttt{[CLS]}}, h_1, \ldots, h_n\}$, where $h_{\texttt{[CLS]}} \in \mathbb{R}^{768}$ represents the document-level embedding. The model is trained using binary cross-entropy to accommodate the XMTC setting.

3.2 Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denote a training corpus where x_i is a biomedical article and $y_i \in \{0,1\}^C$ is its multi-label indicator vector over a label set $\mathcal{L}$ of size C, where $y_{i\ell} = 1$ indicates that label ℓ is assigned to document i. Given a pretrained encoder f_θ, the prediction pipeline composes two functions: $x_i \xrightarrow{f_\theta} h_i \xrightarrow{g_\psi} \hat{y}_i$, where $f_\theta : \mathcal{X} \to \mathbb{R}^d$ maps a document to a contextual representation and $g_\psi : \mathbb{R}^d \to [0,1]^C$ is a linear classification head that outputs label probabilities. The model parameters (θ, ψ) are optimized by minimizing the binary cross-entropy loss.

Overview of SPUR-XMC. Label distributions in large-scale biomedical literature are long-tailed: a few concepts are frequent, while most appear sparsely. This imbalance biases the training toward head labels and causes the PLMs to capture non-informative co-occurrences rather than true semantics. To address this, we propose SPUR-XMC that (1) identifies the geometric directions associated with spurious correlations and (2) attenuates their influence at the document-label level. Moreover, we implement this idea through an add-on enhancer applied on top of a PLM, allowing the model to mitigate frequency-driven spuriousness while preserving the semantic structure of the representation space. Figure 2 shows an overview of the proposed approach.

3.2.1 Frequency-invariant Spurious Estimate

In biomedical literature, estimating a spurious direction independently for each label leads to unreliable estimates for rare labels since limited observations provide insufficient evidence to distinguish genuine semantics from incidental co-occurrences. To address this, we propose a frequency-invariant mean estimator that identifies spurious directions while treating labels uniformly. Standard debiasing methods estimate directions via class-conditional mean shifts between documents with and without label ℓ (Belrose et al., 2023), but this mixes spurious and semantic signals by treating all documents equally. To minimize these effects, we separate abstract–label pairs into two groups using domain knowledge from UMLS (Bodenreider, 2004). For each document i containing label ℓ, the knowledge graph provides a binary indicator $c_{i\ell} \in \{0,1\}$, where $c_{i\ell} = 1$ and $c_{i\ell} = 0$ denote spurious and semantic associations, respectively. This partitions co-occurring documents into two sets: $\mathcal{I}_\ell^{\text{spur}} = \{i : c_{i\ell} = 1\}$ and $\mathcal{I}_\ell^{\text{clean}} = \{i : c_{i\ell} = 0\}$. The corresponding mean representations are:

$$\hat{\mu}_\ell^{\text{spur}} = \frac{1}{|\mathcal{I}_\ell^{\text{spur}}|} \sum_{i \in \mathcal{I}_\ell^{\text{spur}}} h_i, \quad \hat{\mu}_\ell^{\text{clean}} = \frac{1}{|\mathcal{I}_\ell^{\text{clean}}|} \sum_{i \in \mathcal{I}_\ell^{\text{clean}}} h_i, \quad (1)$$

where $h_i \in \mathbb{R}^d$ is the pretrained embedding of document i.

The estimated spurious direction is

$$\hat{v}_\ell = \frac{\hat{\mu}_\ell^{\text{spur}} - \hat{\mu}_\ell^{\text{clean}}}{\|\hat{\mu}_\ell^{\text{spur}} - \hat{\mu}_\ell^{\text{clean}}\|_2}. \quad (2)$$

For each label $\ell \in \mathcal{L}$, we assume $\|\mu_\ell^{\text{spur}} - \mu_\ell^{\text{clean}}\|_2 > 0$, so that the population spurious direction v_ℓ is well-defined.

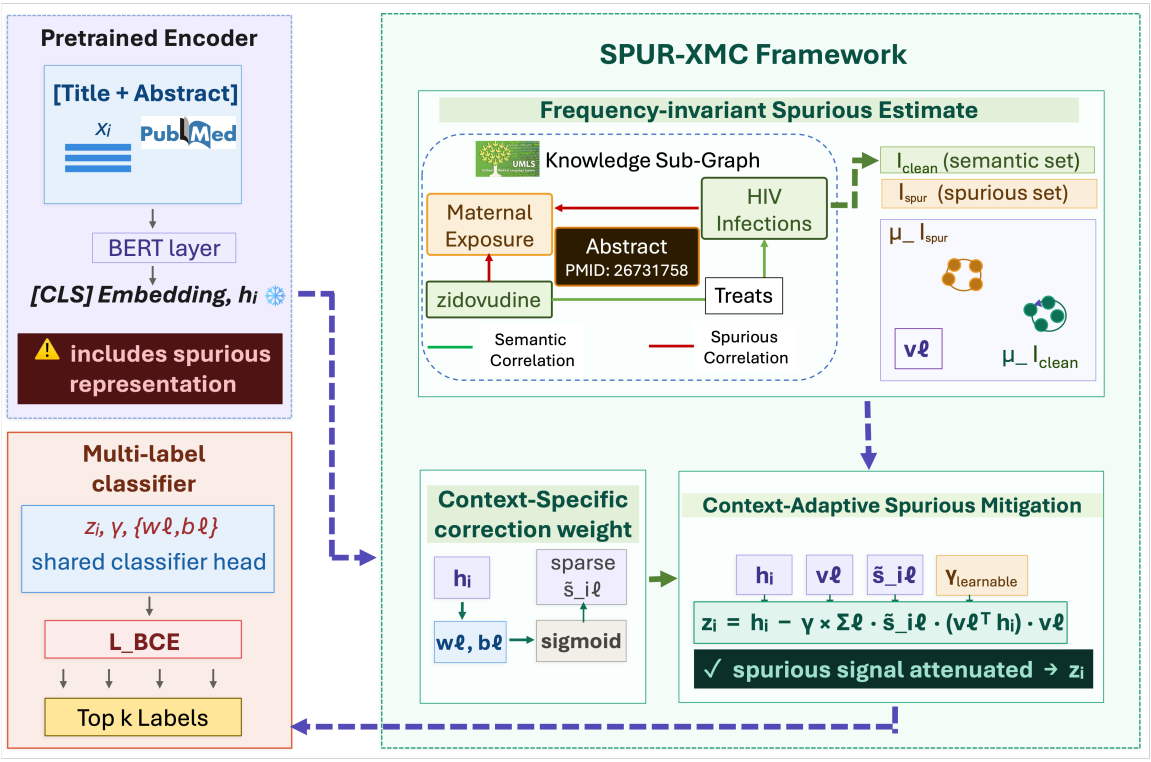


Figure 2: Overview of SPUR-XMC. A pretrained encoder generates biased representations, from which frequency-invariant spurious directions are estimated using biomedical ontologies (Bodenreider, 2004). A context-specific coefficient then modulates these directions to produce corrected representations for classification.

Consider two MeSH labels: *Stroke*, appearing in approximately 11% of documents, and *Paraneoplastic Syndromes, Nervous System*, appearing in fewer than 0.5%. A naive mean-shift estimator directly computes $\hat{\mu}_\ell^{\text{spur}} - \hat{\mu}_\ell^{\text{clean}}$, yielding a stable estimate for *Stroke* but an unstable near-zero estimate for the rare label due to limited observations. Such estimates under-correct tail labels. Normalization in Equation (2) preserves only directional information, making the estimator frequency-invariant and ensuring equal-norm spurious directions across labels. Proposition 3.1 formalizes this property.

Proposition 3.1 (Frequency-Invariant Direction Estimation)**.** *Let $\hat{v}_\ell$ be defined in Equation (2). Then $\|\hat{v}_\ell\|_2 = 1$ for all $\ell \in \mathcal{L}$, independent of sample size. Moreover, as $|\mathcal{I}_\ell^{spur}|, |\mathcal{I}_\ell^{clean}| \to \infty$, $\hat{v}_\ell \to v_\ell$ almost surely, where v_ℓ denotes the population spurious direction. (See Appendix A.2 for proof.)*

Interpretation. The vector $\hat{v}_\ell$ points from the mean of semantically grounded documents toward spuriously associated documents, capturing the direction of frequency-driven bias. Crucially, normalization decouples the estimated direction from label frequency: $\hat{v}_\ell$ has unit norm regardless of how many documents carry label ℓ, ensuring that tail labels receive corrections of the same geometric magnitude as head labels. Without this property, naive mean-shift estimators produce systematically smaller correction vectors for rare labels due to lower-count estimates, effectively leaving the most correction-critical labels underserved.

3.2.2 Context-Specific Spurious Correction

The estimated spurious directions capture corpus-level correlations, but their spuriousness varies with semantic context derived from co-occurring concepts in the knowledge graph. Applying uniform mitigation across all instances of a label can therefore remove meaningful signal while addressing spurious bias. As such, we introduce a document and label-specific correction weight $\tilde{s}_{i\ell} \in (0, 1)$ that determines the extent to which the direction v_ℓ is removed from h_i.We derive $\tilde{s}_{i\ell}$ as follows:

$$\tilde{s}_{i\ell} = \sigma(w_\ell^\top h_i + b_\ell), \tag{3}$$

where w_ℓ and b_ℓ are the weight and bias parameters of the classification head for label ℓ, and $\sigma(\cdot)$ denotes the sigmoid function.A natural concern is that $\tilde{s} * i\ell$ in Equation (3) is derived from the same classifier weights $\mathbf{w} * \ell$ that generate the pre-correction score, potentially creating a feedback loop. How-

ever, the gate is computed from the *uncorrected* embedding $\mathbf{h}_i$ and therefore does not depend on the corrected representation $\mathbf{z}_i$ within the same forward pass; any feedback occurs only across training steps. Moreover, because γ is learned through a softplus reparameterization and the correction is additive, the update magnitude remains controlled.

The learned weights $\mathbf{w} * \ell$ thus implicitly capture how the KG-derived indicator $c * i\ell$ separates spurious from semantically grounded co-occurrences. Accordingly, $\tilde{s} * i\ell$ scales the correction based on document-specific activation, applying stronger removal when an association is likely spurious and weaker removal when it is semantically grounded. Importantly, it modulates only the correction *magnitude*; the direction $\hat{\mathbf{v}} * \ell$ is fixed offline by the KG-guided partitioning in Section 3.2.1.

Although frequency-invariant estimation improves comparability across labels, alignment between $\hat{\mathbf{v}}_\ell$ and classifier weights may still vary, particularly for rare labels with limited observations. Hence, uniform correction can affect classification margins unevenly, motivating document- and label-specific correction weights.

Proposition 3.2 (Suboptimality of Uniform Direction Removal)**.** *Let $r_\ell(h_i) = h_i - \alpha_\ell(v_\ell^\top h_i)v_\ell$ be a uniform scalar-weighted subtraction, where $\alpha_\ell > 0$ is a fixed scalar independent of document and label context. When $w_\ell^\top v_\ell > 0$ and label ℓ is genuinely predictive of document x_i, uniformly applying r_ℓ reduces the expected classification margin on genuinely associated documents. When $w_\ell^\top v_\ell < 0$, uniform application of r_ℓ increases the classification score for both spurious and semantically grounded associations. Since the sign of $w_\ell^\top v_\ell$ varies across the label space, no fixed α_ℓ can simultaneously prevent margin reduction in the first case and suppress score inflation in the second.(See Appendix A.3 for proof.)*

Interpretation. Uniformly removing v_ℓ suppresses not only spurious correlations but also genuine semantic signal aligned with that direction, while for labels where the classification head opposes the spurious direction, it can artificially increase classification scores irrespective of whether the association is semantic or spurious. Since the sign of $w_\ell^\top v_\ell$ varies across the label space, a fixed scalar α_ℓ cannot provide consistent mitigation across all $\ell \in \mathcal{L}$. Empirical analysis in A.9.3 confirms that both alignment conditions co-exist across domains, motivating the context-adaptive mitigation operation introduced in Section 3.2.3.

3.2.3 Context-Adaptive Spurious Mitigation

Having established that spurious correction must operate at the instance level, we now describe how we attenuate the spurious bias from the representation. For a document x_i with representation h_i, we adjust the representation by subtracting the components aligned with the spurious directions, scaled by the instance label weights introduced in section 3.2.2. The corrected representation is computed as:

$$z_i = h_i - \gamma \sum_{\ell \in \mathcal{L}^*} \tilde{s}_{i\ell}(v_\ell^\top h_i)v_\ell, \tag{4}$$

where $v_\ell^\top h_i$ measures the projection of h_i onto the spurious direction associated with label ℓ, $\tilde{s}_{i\ell}$ is the instance-specific correction weight defined in Equation (3), and $\gamma > 0$ is a correction scale shared across all labels and documents. We implement this correction through a translation operation rather than a hard projection. Translation adjusts representations along the spurious directions while preserving the overall geometry of the latent space, whereas projection can distort the representation distribution (Lu et al., 2025; Zhou and Zhu, 2025). In addition, translation allows the correction magnitude to be continuously modulated by the instance label weights $\tilde{s}_{i\ell}$, which would not be possible with a global projection. To prevent corrections from accumulating across many correlated directions, we retain only the top-k correction weights per document and set the remaining weights to zero, where k is chosen as the mean number of labels per document in the training set. The correction strength γ is parameterized as $\gamma = \mathrm{softplus}(\hat{\gamma})$, where $\hat{\gamma}$ is a learnable scalar, ensuring $\gamma > 0$ while allowing the model to adapt the correction magnitude during training. Ablation study of the γ and k design choices is reported in Appendix A.9.

Training Objective. The model is trained by minimizing the binary cross-entropy loss:

$$\mathcal{L}_{\text{task}} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sum_{\ell \in \mathcal{L}} \ell_{\text{bce}}(y_{i\ell}, \hat{y}_{i\ell})\right], \tag{5}$$

$$\ell_{\text{bce}}(y_{i\ell}, \hat{y}_{i\ell}) = -y_{i\ell}\log\hat{y}_{i\ell} - (1 - y_{i\ell})\log(1 - \hat{y}_{i\ell}), \tag{6}$$

and the predicted probability for label ℓ is $\hat{y}_{i\ell} = \sigma(w_\ell^\top z_i + b_\ell)$. Here z_i denotes the debiased representation obtained from Equation (4), and $y_{i\ell} \in \{0, 1\}$ is the ground-truth indicator for label ℓ. Gradients propagate through z_i into both the classification head parameters $\{W_c, b_c\}$ and the correction

Dataset	BioASQ	Embryology	Immunology	Neurology
Training	1,064,845	41,399	13,809	11,052
Val.	152,122	6,786	4,900	3,447
Test	304,244	3,097	2,390	1,550
Labels	28,658	14,116	9,313	2,567

Table 1: Statistics on biomedical literature datasets.

parameter $\log_\gamma$, allowing the classifier and the correction mechanism to co-adapt during fine-tuning. The BERT encoder f_θ remains frozen throughout training, with only the classification head parameters $\{\mathbf{W}_c, \mathbf{b}_c\}$ and correction scale $\log \gamma$ optimized via gradient descent. This design choice constrains spurious mitigation to operate entirely within the pretrained representation space without modifying the encoder architecture.

4 Experiments

Datasets. Table 1 summarizes the benchmark biomedical literature dataset statistics spanning three domains. We choose these datasets as they are widely used in prior studies on biomedical literature understanding (Shariatmadari et al., 2025a; Tyagin and Safro, 2024; Shariatmadari et al., 2025b). To evaluate scalability, we additionally included a BioASQ dataset comprising over 1.5 million articles, derived from the BioASQ Task A corpus (Tsatsaronis et al., 2015). As shown in Figure 4 in A.1, the label space is highly skewed and follows a long-tail distribution, making them well-suited for evaluating rare-label generalization. The datasets and source code used to implement the SPUR-XMC are publicly available at: `https://github.com/ratri1310/spur_xmc`.

Experimental Setup. Our experiments are conducted on a machine with 4 NVIDIA A30 GPUs with 24GiB of memory using the PyTorch framework (Paszke et al., 2019). For fair comparison, we use identical data splits and evaluation protocols across all methods. We train the model for 30 epochs with a batch size of 32 and apply a warmup schedule for the first 5 epochs. The learning rate is set to 1×10^{-3} enabling mixed-precision training. Each experiment is repeated 10 times, and the mean±standard deviation of the results is reported.

Baseline Algorithms. We evaluate SPUR-XMC using widely adopted biomedical pretrained language models, including PubMedBERT (Gu et al., 2021), BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), MedicalBERT (Reddy et al., 2025), ClinicalModernBERT (Lee et al., 2025), SapBERT (Liu et al., 2021), and SciBERT (Beltagy et al., 2019). All models are fine-tuned for XMTC using a linear classification head over the `[CLS]` representation with binary cross-entropy loss. We additionally compare against some state-of-the-art approaches for XMTC BERTMeSH (You et al., 2021), PECOS (Yu et al., 2022), and Fast Multi-Resolution Transformer (Zhang et al., 2021). Additional model details and results are provided in A.4.

Evaluation Metrics. Following prior work in XMTC literature (Jain et al., 2019; You et al., 2019b), we evaluate model performance using Precision@k (P@k) and normalized discounted cumulative gain@k (nDCG@k) for $k \in \{3, 5\}$. Both P@k and nDCG@k emphasize performance on the highest-ranked predictions, which are most relevant in large label spaces. Details of the metrics are shared in the Appendix A.5.

4.1 Results

Table 2 reports the performance of SPUR-XMC across four biomedical datasets. The proposed approach consistently improves P@3, P@5, nDCG@3, and nDCG@5 across backbone encoders and datasets, indicating better recovery and ranking of relevant labels. Consistent gains over encoder-specific baselines and extreme classification methods (XR-Linear (Yu et al., 2022) and XR-Transformer (Zhang et al., 2021)) demonstrate that the improvements generalize beyond specific architectures. While Neurology shows comparatively smaller gains due to stronger baseline performance, larger improvements in Immunology, Embryology, and BioASQ suggest greater benefits under more diverse label distributions. Statistical significance (denoted †) is assessed via a paired t-test ($p < 0.05$) comparing proposed versus baseline over the 10 independent runs for each backbone.

4.2 Performance on Tail Labels

Table 3 reports performance on tail labels, defined as labels appearing in fewer than 5% of documents within each dataset. SPUR-XMC consistently improves P@3, P@5, nDCG@3, and nDCG@5 across backbone encoders and datasets, demonstrating stronger retrieval and ranking performance for low-frequency concepts. The improvements are particularly evident for Immunology, Embryology, and BioASQ, where larger label diversity

Base Model	Method	Immunology				Embryology			
		P@3	P@5	nDCG@3	nDCG@5	P@3	P@5	nDCG@3	nDCG@5
PubMedBERT	Baseline	0.578 ± 0.025	0.467 ± 0.019	0.615 ± 0.013	0.548 ± 0.040	0.732 ± 0.011	0.620 ± 0.011	0.780 ± 0.037	0.669 ± 0.030
	Proposed	**0.626**†± 0.019	**0.523**†± 0.030	**0.672**†± 0.023	**0.590**†± 0.031	**0.774**†± 0.017	**0.655**†± 0.018	**0.813** ± 0.034	**0.722**†± 0.030
BioBERT	Baseline	0.585 ± 0.016	0.487 ± 0.017	0.613 ± 0.013	0.542 ± 0.015	0.745 ± 0.031	0.620 ± 0.035	0.764 ± 0.030	0.668 ± 0.025
	Proposed	**0.626**†± 0.022	**0.523**†± 0.021	**0.672**†± 0.030	**0.591**†± 0.038	**0.782**†± 0.026	**0.660**†± 0.034	**0.821**†± 0.020	**0.727**†± 0.025
BERTMeSH	Baseline	0.627 ±0.014	0.516 ±0.021	0.664 ±0.019	0.598 ±0.023	0.786 ±0.013	0.672 ±0.017	0.833 ±0.021	0.721 ±0.024
	Proposed	**0.674**†±0.012	**0.574**†±0.018	**0.722**†±0.015	**0.641**†±0.019	**0.828**†±0.014	**0.707**†±0.019	**0.866**†±0.018	**0.774**†±0.021
ClinicalModernBERT	Baseline	0.639 ±0.021	0.546 ±0.019	0.681 ±0.018	0.608 ±0.023	0.785 ±0.020	0.660 ±0.019	0.814 ±0.024	0.730 ±0.026
	Proposed	**0.718**†±0.017	**0.612**†±0.015	**0.769**†±0.021	**0.686**†±0.018	**0.820**†±0.018	**0.715**†±0.017	**0.865**†±0.020	**0.778**†±0.021
	XR-Linear	0.671 ±0.020	0.564 ±0.018	0.718 ±0.022	0.639 ±0.019	0.781 ±0.019	0.594 ±0.017	0.728 ±0.021	0.627 ±0.019
	XR-Transformer	0.704 ±0.018	0.597 ±0.016	0.752 ±0.020	0.672 ±0.017	0.814 ±0.017	0.627 ±0.015	0.762 ±0.019	0.660 ±0.017
	XR-Tx+SPUR-XMC	**0.731**†±0.016	**0.624**†±0.014	**0.779**†±0.018	**0.699**†±0.016	**0.843**†±0.015	**0.656**†±0.013	**0.791**†±0.017	**0.689**†±0.015

Base Model	Method	Neurology				BioASQ			
		P@3	P@5	nDCG@3	nDCG@5	P@3	P@5	nDCG@3	nDCG@5
PubMedBERT	Baseline	0.747 ±0.032	0.623 ±0.011	0.800 ±0.021	0.707 ±0.011	0.729 ±0.017	0.617 ±0.014	0.778 ±0.022	0.651 ±0.018
	Proposed	**0.766**†±0.038	**0.641**†±0.030	**0.811** ±0.038	**0.721**†±0.028	**0.781**†±0.015	**0.655**†±0.017	**0.821**†±0.019	**0.712**†±0.020
BioBERT	Baseline	0.764 ±0.031	0.623 ±0.026	0.800 ±0.033	0.708 ±0.029	0.749 ±0.018	0.627 ±0.019	0.768 ±0.021	0.581 ±0.017
	Proposed	**0.767**†±0.022	**0.644**†±0.015	**0.813** ±0.017	**0.724**†±0.034	**0.783**†±0.014	**0.662**†±0.018	**0.834**†±0.018	**0.720**†±0.019
BERTMeSH	Baseline	0.747 ±0.019	0.623 ±0.018	0.800 ±0.021	0.707 ±0.023	0.781 ±0.014	0.667 ±0.016	0.829 ±0.019	0.704 ±0.018
	Proposed	**0.766**†±0.017	**0.641**†±0.019	**0.811**†±0.020	**0.721**†±0.021	**0.833**†±0.013	**0.705**†±0.017	**0.872**†±0.018	**0.762**†±0.019
ClinicalModernBERT	Baseline	0.850 ±0.020	0.705 ±0.019	0.884 ±0.023	0.800 ±0.025	0.758 ±0.017	0.635 ±0.016	0.792 ±0.021	0.668 ±0.018
	Proposed	**0.865**†±0.017	**0.735**†±0.016	**0.905**†±0.019	**0.815**†±0.020	**0.806**†±0.014	**0.689**†±0.017	**0.848**†±0.019	**0.748**†±0.020
	XR-Linear	0.654 ±0.021	0.559 ±0.017	0.698 ±0.022	0.623 ±0.020	0.791 ±0.018	0.649 ±0.016	0.755 ±0.019	0.634 ±0.018
	XR-Transformer	0.687 ±0.018	0.592 ±0.015	0.731 ±0.019	0.656 ±0.018	0.824 ±0.016	0.682 ±0.014	0.788 ±0.017	0.667 ±0.016
	XR-Tx+SPUR-XMC	**0.712**†±0.016	**0.617**†±0.013	**0.756**†±0.017	**0.681**†±0.016	**0.853**†±0.014	**0.711**†±0.012	**0.817**†±0.016	**0.696**†±0.014

Table 2: Performance comparison across four datasets. Results are reported as mean ± standard deviation; bold values denote the best performance and †indicates statistically significant improvement.

Base Model	Method	Immunology				Embryology			
		P@3	P@5	nDCG@3	nDCG@5	P@3	P@5	nDCG@3	nDCG@5
PubMedBERT	Baseline	0.447 ± 0.017	0.361 ± 0.029	0.475 ± 0.021	0.421 ± 0.018	0.585 ± 0.026	0.508 ± 0.018	0.616 ± 0.015	0.533 ± 0.022
	Proposed	**0.521**†± 0.028	**0.434**†± 0.011	**0.557**†± 0.019	**0.488**†± 0.026	**0.650**†± 0.014	**0.557**†± 0.027	**0.684**†± 0.024	**0.607**†± 0.019
BioBERT	Baseline	0.452 ± 0.014	0.376 ± 0.023	0.474 ± 0.027	0.419 ± 0.012	0.584 ± 0.016	0.508 ± 0.023	0.599 ± 0.030	0.520 ± 0.017
	Proposed	**0.538**†± 0.025	**0.451**†± 0.020	**0.581**†± 0.013	**0.508**†± 0.024	**0.659**†± 0.025	**0.559**†± 0.013	**0.700**†± 0.018	**0.627**†± 0.026
BERTMeSH	Baseline	0.502 ± 0.012	0.411 ± 0.021	0.529 ± 0.028	0.478 ± 0.015	0.630 ± 0.011	0.520 ± 0.025	0.657 ± 0.023	0.567 ± 0.014
	Proposed	**0.579**†± 0.024	**0.493**†± 0.019	**0.623**†± 0.013	**0.551**†± 0.022	**0.710**†± 0.026	**0.591**†± 0.012	**0.750**†± 0.015	**0.646**†± 0.024
ClinicalModernBERT	Baseline	0.514 ± 0.028	0.439 ± 0.016	0.548 ± 0.019	0.489 ± 0.030	0.612 ± 0.015	0.533 ± 0.030	0.653 ± 0.022	0.580 ± 0.018
	Proposed	**0.599**†± 0.023	**0.513**†± 0.011	**0.644**†± 0.014	**0.575**†± 0.025	**0.703**†± 0.013	**0.599**†± 0.021	**0.721**†± 0.028	**0.669**†± 0.012
	XR-Linear	0.529 ± 0.017	0.447 ± 0.029	0.570 ± 0.022	0.507 ± 0.016	0.626 ± 0.025	0.475 ± 0.017	0.584 ± 0.011	0.493 ± 0.026
	XR-Transformer	0.549 ± 0.014	0.463 ± 0.020	0.587 ± 0.027	0.521 ± 0.011	0.649 ± 0.029	0.490 ± 0.016	0.619 ± 0.018	0.518 ± 0.030
	XR-Tx+SPUR-XMC	**0.630**†± 0.015	**0.537**†± 0.028	**0.671**†± 0.026	**0.599**†± 0.019	**0.711**†± 0.027	**0.547**†± 0.015	**0.673**†± 0.019	**0.581**†± 0.028

Base Model	Method	Neurology				BioASQ			
		P@3	P@5	nDCG@3	nDCG@5	P@3	P@5	nDCG@3	nDCG@5
PubMedBERT	Baseline	0.610 ± 0.012	0.483 ± 0.024	0.656 ± 0.029	0.575 ± 0.013	0.577 ± 0.028	0.484 ± 0.017	0.600 ± 0.011	0.508 ± 0.026
	Proposed	**0.659**†± 0.025	**0.546**†± 0.011	**0.697**†± 0.014	**0.627**†± 0.023	**0.676**†± 0.029	**0.561**†± 0.015	**0.703**†± 0.018	**0.604**†± 0.030
BioBERT	Baseline	0.607 ± 0.016	0.506 ± 0.028	0.632 ± 0.022	0.547 ± 0.019	0.608 ± 0.027	0.491 ± 0.016	0.602 ± 0.019	0.469 ± 0.025
	Proposed	**0.653**†± 0.027	**0.542**†± 0.018	**0.691**†± 0.011	**0.621**†± 0.024	**0.675**†± 0.011	**0.575**†± 0.028	**0.710**†± 0.026	**0.607**†± 0.017
BERTMeSH	Baseline	0.589 ± 0.029	0.497 ± 0.013	0.651 ± 0.016	0.566 ± 0.022	0.628 ± 0.024	0.531 ± 0.013	0.661 ± 0.015	0.549 ± 0.022
	Proposed	**0.640**†± 0.011	**0.542**†± 0.026	**0.684**†± 0.024	**0.607**†± 0.015	**0.713**†± 0.016	**0.589**†± 0.021	**0.724**†± 0.030	**0.643**†± 0.019
ClinicalModernBERT	Baseline	0.657 ± 0.028	0.570 ± 0.018	0.693 ± 0.011	0.634 ± 0.029	0.585 ± 0.025	0.504 ± 0.018	0.636 ± 0.011	0.516 ± 0.028
	Proposed	**0.719**†± 0.013	**0.612**†± 0.026	**0.759**†± 0.021	**0.680**†± 0.014	**0.694**†± 0.014	**0.582**†± 0.026	**0.705**†± 0.024	**0.632**†± 0.013
	XR-Linear	0.505 ± 0.012	0.454 ± 0.023	0.554 ± 0.027	0.503 ± 0.015	0.620 ± 0.019	0.526 ± 0.030	0.613 ± 0.027	0.503 ± 0.016
	XR-Transformer	0.529 ± 0.025	0.461 ± 0.011	0.572 ± 0.014	0.514 ± 0.024	0.662 ± 0.028	0.556 ± 0.015	0.630 ± 0.018	0.524 ± 0.025
	XR-Tx+SPUR-XMC	**0.595**†± 0.016	**0.519**†± 0.029	**0.640**†± 0.022	**0.584**†± 0.017	**0.715**†± 0.013	**0.609**†± 0.029	**0.700**†± 0.023	**0.600**†± 0.014

Table 3: Performance comparison across four datasets on tail labels. Results are reported as mean ± standard deviation; bold values denote the best performance and †indicates statistically significant improvement.

and long-tailed distributions introduce greater challenges for representation learning. The consistent gains across both pretrained language models and extreme classification baselines suggest that the proposed mitigation strategy effectively reduces frequency-driven bias while preserving meaningful semantic information for rare labels.

4.3 Qualitative Analysis

Table 4 shows articles with top-5 predictions from the baseline and proposed model, where rare labels are highlighted, and check/cross marks denote correct/incorrect predictions. The BioBERT baseline encoder (Lee et al., 2020) tends to prioritize frequent labels (e.g., *Animals*), while the proposed

PMID	Abstract	Ground Labels	Baseline Top-5	Proposed Top-5
Case 1: Rare Label Recovery				
28847650	Embryonic hypoxia and kidney development mediated by...	Anemia; Cell Hypoxia; Hypoxia-Inducible Factor; Kidney Diseases; Hydroxylation	Animals (✗); Cell Hypoxia (✓); Enzyme Inhibitors (✗); Kidney (✗); Stromal Cells (✗)	Animals (✗); Cell Hypoxia (✓); Kidney (✗); Hypoxia-Inducible Factor (✓); Hydroxylation (✓); Kidney Diseases (✓)
Insight: The baseline predicts broad labels, whereas the proposed model identifies rare hypoxia-related concepts				
Case 2: Improved Ranking of Informative Labels				
21565280	Isoprostanes induce vascular smooth muscle contraction in chick embryos.	Chick Embryo; Myography; Thromboxane Receptor; Vasoconstrictor Agents; Isoprostanes	Animals (✗); Chick Embryo (✓); Isoprostanes (✓); Muscle Contraction (✗); Oxidative Stress (✗)	Chick Embryo (✓); Isoprostanes (✓); Myography (✓); Thromboxane Receptor (✓); Vasoconstrictor Agents (✓)
Insight: The baseline prioritizes generic labels (e.g., Animals), while the proposed model elevates document-specific concepts.				

Table 4: Qualitative examples of biomedical abstracts. Rare labels (green); The baseline favors frequent labels,the proposed method retrieves document-specific rare labels aligned with the ground truth.

method retrieves semantically aligned labels, including rare concepts missed by the baseline. It also ranks informative labels higher. These examples demonstrate that context-adaptive mitigation reduces frequency-driven bias and promotes semantically meaningful biomedical concepts. Figure 3 shows t-SNE visualizations of document embeddings on the Immunology dataset. The pretrained BioBERT embeddings exhibit overlap across concepts, while the proposed method produces better-separated clusters, indicating improved alignment.

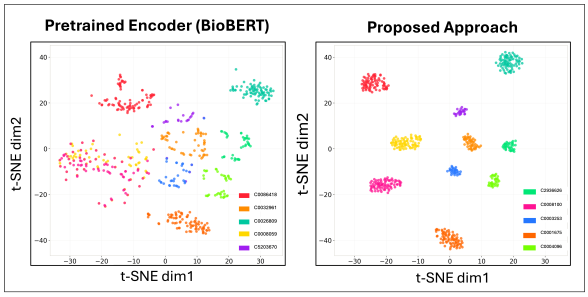


Figure 3: t-SNE visualization of label–document interactions for the Immunology dataset. The pretrained encoder (left) shows overlapping clusters; the proposed method (right) produces clearly separated clusters.

4.4 Ablation Study

Table 5 evaluates the contribution of individual components in the proposed spurious mitigation framework across four biomedical domains, including BioASQ. Each additional component consistently improves nDCG@3, nDCG@5, and mAP, with full SPUR-XMC achieving the best overall performance. These findings suggest that frequency-invariant normalization and adaptive scaling enhance instance-level mitigation by more effectively modeling label and context-specific spurious correlations. Further ablation varying top-k parameter and scalar versus per-label correction parameters are discussed in A.9.

Dataset	Method	nDCG@3↑	nDCG@5↑	mAP↑
Immunology	w/o Instance Weight	0.583	0.512	0.268
	w/o Freq-Inv. Norm	0.618	0.541	0.271
	w/o γ	0.648	0.568	0.315
	SPUR-XMC	**0.672**	**0.591**	**0.341**
Embryology	w/o Instance Weight	0.736	0.638	0.382
	w/o Freq-Inv. Norm	0.759	0.652	0.416
	w/o γ	0.797	0.683	0.425
	SPUR-XMC	**0.821**	**0.727**	**0.459**
Neurology	w/o Instance Weight	0.770	0.676	0.398
	w/o Freq-Inv. Norm	0.784	0.683	0.433
	w/o γ	0.795	0.702	0.456
	SPUR-XMC	**0.813**	**0.724**	**0.466**
BioASQ	w/o Instance Weight	0.716	0.624	0.393
	w/o Freq-Inv. Norm	0.728	0.648	0.401
	w/o γ	0.793	0.692	0.422
	SPUR-XMC	**0.834**	**0.720**	**0.492**

Table 5: Ablation study comparing spurious correlation mitigation components against full SPUR-XMC.

5 Conclusion

In this work, we proposed **SPUR-XMC**, a context-adaptive spurious correlation mitigation approach for XMTC of biomedical literature. Specifically, SPUR-XMC selectively attenuates spurious influences according to the context of the input article, thus improving robustness under long-tail distributions. Moreover, the modular nature of SPUR-XMC enables its flexible integration with various off-the-shelf backbone encoders designed for XMTC of biomedical literature.

Limitations

While SPUR-XMC advances XMTC for biomedical literature classification, several limitations remain. First, it has not been evaluated on broader tasks such as citation prediction, literature search, and summarization, which could better assess its robustness and scalability. Second, our study focuses primarily on BERT-based architectures; exploring alternatives such as cross-encoders and late-interaction models (Humeau et al., 2019) would be valuable. Finally, SPUR-XMC currently doesn't offer interpretable explanations, which are important for deployment in high-stakes biomedical applications, particularly with generative AI.

Acknowledgment

This work is supported in part by the U.S. National Science Foundation (NSF) and the National Institutes of Health (NIH) under grants SCH-2500344 and R01LM014012-01A1. Any recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NIH or NSF.

References

Emily Alsentzer, John Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3615–3620.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and 1 others. 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S Dhillon. 2020. Taming pretrained transformers for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3163–3171.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Nilesh Gupta, Sakina Bohra, Yashoteja Prabhu, Saurabh Purohit, and Manik Varma. 2021. Generalized zero-shot extreme multi-label learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 527–535.

Hyeonggeun Han, Sehwan Kim, Hyungjun Joo, Sangwoo Hong, and Jungwoo Lee. 2024. Mitigating spurious correlations via disagreement probability. *Advances in Neural Information Processing Systems*, 37:74363–74382.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*.

Himanshu Jain, Venkatesh Balasubramanian, Bhanu Chunduri, and Manik Varma. 2019. Slice: Scalable linear extreme classifiers trained on 100 million labels for related searches. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 528–536.

Himanshu Jain, Yashoteja Prabhu, and Manik Varma. 2016. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 935–944.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Kamal Raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. Bioelectra: pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th workshop on biomedical language processing*, pages 143–154.

Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. 2023. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Simon A Lee, Anthony Wu, and Jeffrey N Chiang. 2025. Clinical modernbert: An efficient and long context encoder for biomedical text. *arXiv preprint arXiv:2504.03964*.

Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The unified medical language system. *Yearbook of medical informatics*, 2(01):41–51.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4228–4238.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 115–124.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shenyu Lu, Junyi Chai, and Xiaoqian Wang. 2025. Mitigating spurious correlations in zero-shot multimodal models. In *The Thirteenth International Conference on Learning Representations*.

Zhiyong Lu. 2011. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011:baq036.

Anshul Mittal, Noveen Sachdeva, Sheshansh Agrawal, Sumeet Agarwal, Purushottam Kar, and Manik Varma. 2021. Eclare: Extreme classification with label graph correlations. In *Proceedings of the Web Conference 2021*, pages 3721–3732.

Ratri Mukherjee, Shailesh Dahal, and Kishlay Jha. 2025. Disentangled contrastive representation learning for zero-shot biomedical text classification. In *2025 IEEE International Conference on Data Mining (ICDM)*, pages 1425–1434. IEEE.

Ratri Mukherjee and Kishlay Jha. 2024. Context-aware contrastive representation learning for zero-shot biomedical text classification. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3611–3614. IEEE.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

K Sahit Reddy, N Ragavenderan, Ganesh N Naik, Vishalakshi Prabhu, and 1 others. 2025. Medicalbert: enhancing biomedical natural language processing using pretrained bert-based model. *arXiv preprint arXiv:2507.08013*.

Deepak Saini, Arnav Kumar Jain, Kushal Dave, Jian Jiao, Amit Singh, Ruofei Zhang, and Manik Varma. 2021. Galaxc: Graph neural networks with labelwise attention for extreme classification. In *Proceedings of the web conference 2021*, pages 3733–3744.

Amir Hassan Shariatmadari, Sikun Guo, Nathan C Sheffield, Aidong Zhang, and Kishlay Jha. 2025a. Hyhg: A temporal hypergraph contrastive learning framework for biomedical hypothesis generation. In *2025 IEEE International Conference on Data Mining (ICDM)*, pages 703–712. IEEE.

Amir Hassan Shariatmadari, Alireza Jafari, Sikun Guo, Sneha Srinivasan, Nathan C Sheffield, Aidong Zhang, and Kishlay Jha. 2025b. Conceptdrift: leveraging spatial, temporal and semantic evolution of biomedical concepts for hypothesis generation. *Bioinformatics*, 41(11):btaf563.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, and 1 others. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Ilya Tyagin and Ilya Safro. 2024. Dyport: dynamic importance-based biomedical hypothesis generation benchmarking technique. *BMC bioinformatics*, 25(1):213.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022. Identifying and mitigating spurious correlations for improving robustness in NLP models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1719–1729, Seattle, United States. Association for Computational Linguistics.

Tong Wei, Wei-Wei Tu, Yu-Feng Li, and Guo-Ping Yang. 2021. Towards robust prediction on tail labels. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 1812–1820.

Xiwei Xuan, Ziquan Deng, Hsuan-Tien Lin, and Kwan-Liu Ma. 2024. Slim: Spuriousness mitigation with minimal human annotations. In *European Conference on Computer Vision*, pages 215–231. Springer.

Ronghui You, Yuxuan Liu, Hiroshi Mamitsuka, and Shanfeng Zhu. 2021. Bertmesh: deep contextual representation learning for large-scale high-performance mesh indexing with full text. *Bioinformatics*, 37(5):684–692.

Ronghui You, Zihan Zhang, Suyang Dai, and Shanfeng Zhu. 2019a. Haxmlnet: Hierarchical attention network for extreme multi-label text classification. *arXiv preprint arXiv:1904.12578*.

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019b. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. *Advances in neural information processing systems*, 32.

Hsiang-Fu Yu, Kai Zhong, Jiong Zhang, Wei-Cheng Chang, and Inderjit S Dhillon. 2022. Pecos: Prediction for enormous and correlated output spaces. *Journal of Machine Learning Research*, 23(98):1–32.

Jiong Zhang, Wei-Cheng Chang, Hsiang-Fu Yu, and Inderjit Dhillon. 2021. Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. *Advances in Neural Information Processing Systems*, 34:7267–7280.

Sendong Zhao, Chang Su, Zhiyong Lu, and Fei Wang. 2021. Recent advances in biomedical literature mining. *Briefings in Bioinformatics*, 22(3):bbaa057.

Guangtao Zheng, Wenqian Ye, and Aidong Zhang. 2024. Learning robust classifiers with self-guided spurious correlation mitigation. *arXiv preprint arXiv:2405.03649*.

Yuqing Zhou and Ziwei Zhu. 2025. Fighting spurious correlations in text classification via a causal learning perspective. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4264–4274.

A Appendices

A.1 Long Tail Distribution in Biomedical Literature

The figure 4 illustrates the rank–frequency distributions of MeSH labels for the Neurology, Immunology, and Embryology datasets on log–log scales, revealing a pronounced long-tailed structure in all three domains. A small number of labels occur frequently, while a large fraction of labels appear only rarely, indicating severe label imbalance and sparsity. Such distributions make supervised learning inadequate, as many clinically meaningful labels lack sufficient training examples, thereby motivating reliable prediction for rare biomedical concepts.

A.2 Proof of Proposition 3.1

Assumption. For each label $\ell \in \mathcal{L}$, we assume

$$\mu_\ell^{\text{spur}} - \mu_\ell^{\text{clean}} \neq 0. \tag{7}$$

Equivalently,

$$\|\mu_\ell^{\text{spur}} - \mu_\ell^{\text{clean}}\|_2 > 0, \tag{8}$$

so that the population spurious direction v_ℓ is well-defined.

Proof. 1. Invariance of the Norm.
By definition in Equation (2),

$$\hat{v}_\ell = \frac{\Delta\hat{\mu}_\ell}{\|\Delta\hat{\mu}_\ell\|_2}, \qquad \Delta\hat{\mu}_\ell = \hat{\mu}_\ell^{\text{spur}} - \hat{\mu}_\ell^{\text{clean}}. \tag{9}$$

For any non-zero vector $u \in \mathbb{R}^d$,

$$\left\| \frac{u}{\|u\|_2} \right\|_2 = \frac{\|u\|_2}{\|u\|_2} = 1. \tag{10}$$

Hence, whenever $\Delta\hat{\mu}_\ell \neq 0$,

$$\|\hat{v}_\ell\|_2 = 1. \tag{11}$$

Therefore, the norm of $\hat{v}_\ell$ is independent of the sample sizes $|\mathcal{I}_\ell^{\text{spur}}|$ and $|\mathcal{I}_\ell^{\text{clean}}|$.

2. Convergence (Consistency).
Let h denote the random variable corresponding to document embeddings, and assume it has finite first moment under both conditional distributions. By the Strong Law of Large Numbers,

$$\hat{\mu}_\ell^{\text{spur}} = \frac{1}{|\mathcal{I}_\ell^{\text{spur}}|} \sum_{i \in \mathcal{I}_\ell^{\text{spur}}} h_i \xrightarrow{\text{a.s.}} \mu_\ell^{\text{spur}}, \tag{12}$$

and

$$\hat{\mu}_\ell^{\text{clean}} = \frac{1}{|\mathcal{I}_\ell^{\text{clean}}|} \sum_{i \in \mathcal{I}_\ell^{\text{clean}}} h_i \xrightarrow{\text{a.s.}} \mu_\ell^{\text{clean}}. \tag{13}$$

By subtraction,

$$\Delta\hat{\mu}_\ell = \hat{\mu}_\ell^{\text{spur}} - \hat{\mu}_\ell^{\text{clean}} \xrightarrow{\text{a.s.}} \mu_\ell^{\text{spur}} - \mu_\ell^{\text{clean}} = \Delta\mu_\ell. \tag{14}$$

By the assumption above, $\Delta\mu_\ell \neq 0$. Therefore, with probability one, $\Delta\hat{\mu}_\ell \neq 0$ eventually for sufficiently large sample size. Since the map

$$f(u) = \frac{u}{\|u\|_2} \tag{15}$$

is continuous on $\mathbb{R}^d \setminus \{0\}$, the Continuous Mapping Theorem gives

$$\hat{v}_\ell = f(\Delta\hat{\mu}_\ell) \xrightarrow{\text{a.s.}} f(\Delta\mu_\ell) = \frac{\Delta\mu_\ell}{\|\Delta\mu_\ell\|_2} = v_\ell. \tag{16}$$

Thus, $\hat{v}_\ell$ is unit-normalized and converges almost surely to the population spurious direction v_ℓ. This formulation is grounded in the linear guardedness framework of Belrose et al. (2023). In particular, their Theorem 3.1 establishes that a representation fails to linearly guard a binary concept if and only if the class-conditional means of that concept differ. Prior work applies this result directly to label prediction, treating the label itself as the binary concept.

In contrast, we apply this principle to the binary concept $c_{i\ell}$ that distinguishes *spurious* from *semantically grounded* document–label associations. Restricting attention to documents with $y_{i\ell} = 1$, the class-conditional means of this concept are given by $\hat{\mu}_\ell^{\text{spur}}$ and $\hat{\mu}_\ell^{\text{clean}}$. According to the linear guardedness result, the direction along which the representation fails to guard $c_{i\ell}$ is precisely the mean difference

$$\hat{v}_\ell \propto \hat{\mu}_\ell^{\text{spur}} - \hat{\mu}_\ell^{\text{clean}}. \tag{17}$$

Thus, while the theoretical guarantee follows from Belrose et al. (2023), our contribution is to redefine the guarded concept itself: rather than guarding the label ℓ, we identify the direction that separates spuriously associated instances from genuine ones. This isolates the geometric footprint of spurious correlations within the positive class and enables targeted mitigation of their effect.

A.3 Proof of Proposition 3.2

Proof. Let

$$m_i = w_\ell^\top h_i \tag{18}$$

denote the pre-sigmoid classification margin for label ℓ on document i. After applying the uniform scalar-weighted subtraction r_ℓ, the representation becomes

$$r_\ell(h_i) = h_i - \alpha_\ell(v_\ell^\top h_i)v_\ell. \tag{19}$$

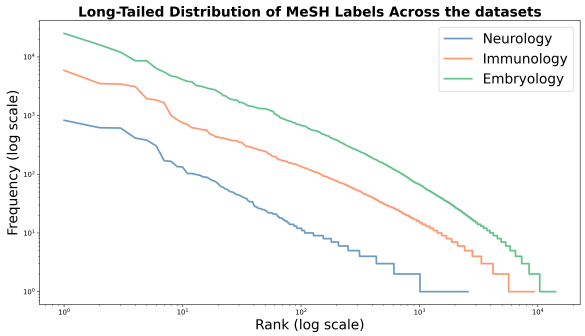


Figure 4: Rank–frequency distributions of MeSH labels in the Neurology, Immunology, and Embryology datasets plotted on log–log scales. All three domains exhibit a pronounced long-tailed structure.

The resulting margin is therefore

$$\begin{aligned} m_i' &= w_\ell^\top r_\ell(h_i) && (20)\\ &= w_\ell^\top \left(h_i - \alpha_\ell(v_\ell^\top h_i)v_\ell\right) && (21)\\ &= w_\ell^\top h_i - \alpha_\ell(w_\ell^\top v_\ell)(v_\ell^\top h_i) && (22)\\ &= m_i - \alpha_\ell(w_\ell^\top v_\ell)(v_\ell^\top h_i). && (23) \end{aligned}$$

Hence the induced change in margin becomes

$$m_i' - m_i = -\alpha_\ell(w_\ell^\top v_\ell)(v_\ell^\top h_i). \tag{24}$$

Case 1: $w_\ell^\top v_\ell > 0$ **(margin reduction).**
For genuinely associated documents, semantic signal aligns with v_ℓ, implying

$$\mathbb{E}[v_\ell^\top h_i \mid c_{i\ell} = 0] > 0. \tag{25}$$

Under $w_\ell^\top v_\ell > 0$ and $\alpha_\ell > 0$,

$$\mathbb{E}[m_i' - m_i \mid c_{i\ell} = 0] = -\alpha_\ell(w_\ell^\top v_\ell)\, \mathbb{E}[v_\ell^\top h_i \mid c_{i\ell} = 0] < 0. \tag{26}$$

Thus uniform removal reduces the expected margin for genuinely associated documents.

Case 2: $w_\ell^\top v_\ell < 0$ **(score inflation).**
For any document satisfying $v_\ell^\top h_i > 0$,

$$m_i' - m_i = -\alpha_\ell \underbrace{(w_\ell^\top v_\ell)}_{<0} \underbrace{(v_\ell^\top h_i)}_{>0} > 0. \tag{27}$$

Thus uniform correction increases classification scores regardless of whether the association is spurious or semantically grounded.

Since both $w_\ell^\top v_\ell > 0$ and $w_\ell^\top v_\ell < 0$ coexist across labels, no fixed α_ℓ can simultaneously avoid

margin reduction in Case 1 and suppress score inflation in Case 2. Therefore, a document- and label-specific correction mechanism is required.

A.4 Baseline Model Details

PubMedBERT. PubMedBERT (Gu et al., 2021) is a domain-specific language model pretrained entirely from scratch on biomedical literature. Unlike earlier biomedical models that continue pretraining from general-domain BERT, PubMedBERT is trained solely on biomedical corpora using a vocabulary constructed from PubMed and PubMed Central articles. The model follows the BERT-base architecture (12 transformer layers, 768 hidden size, 12 attention heads) and is trained using masked language modeling and next sentence prediction objectives. The pretraining corpus consists of approximately 14 million PubMed abstracts together with PubMed Central full-text articles, providing extensive coverage of biomedical terminology. Because it is trained exclusively on biomedical text and uses a biomedical WordPiece vocabulary, PubMedBERT provides strong domain-specific representations and serves as the primary backbone model in our experiments.

BioBERT. BioBERT (Lee et al., 2020) is a biomedical adaptation of the original BERT model obtained through continual pretraining on biomedical text. The model initializes from BERT-base weights trained on Wikipedia and BooksCorpus and is further pretrained on large-scale PubMed abstracts. BioBERT retains the original BERT-base-cased architecture and vocabulary, while the additional pretraining stage exposes the model to domain-specific biomedical terminology. The final version used in this work, BioBERT v1.1, was trained for one million additional steps on PubMed abstracts. BioBERT is one of the earliest and most widely used biomedical language models and represents the classical continual-pretraining paradigm for domain adaptation in biomedical NLP.

ClinicalBERT. ClinicalBERT (Alsentzer et al., 2019) extends biomedical language models into the clinical domain by continuing pretraining on electronic health record (EHR) data. Specifically, the Bio_ClinicalBERT variant is initialized from BioBERT and further trained on clinical notes from the MIMIC-III database. The model retains the BERT-base architecture and vocabulary while learning representations that capture clinical language patterns such as abbreviations and structured documentation styles found in clinical notes. The MIMIC-III corpus used for pretraining contains approximately 880 million words across multiple note categories, including discharge summaries, radiology reports, and nursing notes. ClinicalBERT therefore incorporates both biomedical literature knowledge and clinical documentation patterns, allowing us to evaluate whether exposure to clinical text improves biomedical literature classification.

SapBERT. SapBERT (Liu et al., 2021) introduces a distinct pretraining paradigm based on self-alignment of biomedical entity representations. The model is initialized from PubMedBERT and further trained using a metric learning objective over biomedical synonym pairs derived from the Unified Medical Language System (UMLS). During training, representations of different surface forms referring to the same biomedical concept are encouraged to align in the embedding space using contrastive learning. This self-alignment pretraining stage optimizes representations for biomedical concept normalization and synonym resolution. Because MeSH classification involves predicting biomedical concepts associated with documents, SapBERT provides a strong baseline that explicitly encodes biomedical entity semantics.

SciBERT. SciBERT (Beltagy et al., 2019) is a pretrained language model designed for scientific text. Similar to PubMedBERT, SciBERT is trained from scratch rather than initialized from general-domain BERT. The model uses a custom vocabulary (SciVocab) constructed from scientific publications and is trained on a large corpus of Semantic Scholar papers containing approximately 1.14 million full-text scientific articles. While the majority of the corpus comes from biomedical publications, it also includes computer science literature, providing broader scientific coverage than biomedical-only models. SciBERT therefore serves as a useful baseline for evaluating whether representations trained on general scientific text can generalize to biomedical MeSH classification tasks.

BERTMeSH. BERTMeSH (You et al., 2021) is a multi-label classification framework designed for large-scale MeSH indexing of biomedical literature. The model jointly encodes the title, abstract, and the top-k longest document sections using a shared biomedical encoder, followed by a multi-label attention mechanism that learns label-specific token importance. The at-

tended representations are subsequently processed using a parameter-sharing output layer for efficient large-scale prediction. As the original implementation was not publicly released, we utilize the `WellcomeBertMesh` implementation available through Hugging Face at: https://huggingface.co/Wellcome/WellcomeBertMesh, which employs a PubMedBERT backbone and provides pretrained weights fine-tuned on BioASQ data.

MedicalBERT. MedicalBERT (Reddy et al., 2025) is a domain-specific biomedical language model developed to improve contextual representation learning for medical text mining applications. The model adopts the standard BERT architecture while incorporating a specialized byte-pair encoding (BPE) vocabulary derived from biomedical text to improve representation of rare and complex medical terminology. It is pretrained on a large-scale corpus combining PubMed literature and clinical records from MIMIC-III, enabling the model to capture contextual information from both biomedical and clinical domains.

ClinicalModernBERT. ClinicalModernBERT (Lee et al., 2025) is a modern biomedical language model designed to improve computational efficiency and long-context representation learning. Compared to conventional BERT architectures, it incorporates architectural modifications including rotary positional embeddings (RoPE), SwiGLU activation functions, and optimized attention mechanisms to support improved efficiency and longer context processing. The model is pretrained on a large-scale medical corpus containing clinical narratives and biomedical literature, providing a strong foundation for downstream biomedical natural language processing tasks.

Adapting several global spurious mitigation baselines such as INLP (**?**) and LEACE (Belrose et al., 2023) to XMTC requires a uniform per-label direction subtraction with no instance-level modulation; the *w/o Instance Weight* row in Table 5 already implements this and serves as their functional equivalent. DFR (Kirichenko et al., 2023) requires a balanced, group-annotated subset for head retraining, which is unavailable in this setting and unreliable for tail labels with few positive examples, making it not directly applicable here.

A.5 Details of Evaluation Metrics

XMTC Evaluation Metrics. Precision@k measures the proportion of relevant labels among the top-k predicted labels. While P@k captures how many correct labels appear among the top predictions, it does not consider their relative ranking. To account for ranking quality, we also report normalized discounted cumulative gain (nDCG@k) (Järvelin and Kekäläinen, 2002). These metrics are widely adopted in extreme multi-label classification benchmarks because they emphasize performance on the highest-ranked predictions, which are most relevant in large label spaces where retrieving the most important labels is critical (Jain et al., 2019; You et al., 2019b).

Formally, let $\hat{y}_i^{(k)}$ denote the set of top-k predicted labels for document i, and let y_i denote the ground-truth label set. Precision@k is defined as

$$P@k = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{y}_i^{(k)} \cap y_i|}{k}. \tag{28}$$

While P@k captures how many correct labels appear among the top predictions, it does not consider their relative ranking. To account for ranking quality, we also report normalized discounted cumulative gain (nDCG@k) (Järvelin and Kekäläinen, 2002). This metric assigns higher importance to correct labels appearing earlier in the ranked list:

$$DCG@k = \sum_{j=1}^{k}\frac{\mathrm{rel}_{ij}}{\log_2(j+1)}, \tag{29}$$

where $\mathrm{rel}_{ij} \in \{0, 1\}$ indicates whether the label at rank j for instance i is relevant. The normalized version is computed as

$$nDCG@k = \frac{1}{N}\sum_{i=1}^{N}\frac{DCG@k}{IDCG@k}, \tag{30}$$

where $IDCG@k$ denotes the ideal discounted cumulative gain obtained by a perfect ranking.

Propensity-Scored Evaluation Metrics. Extreme multi-label classification datasets typically exhibit highly long-tailed label distributions, where a small number of head labels appear frequently while a large portion of labels occur only a few times. Standard ranking metrics such as Precision@k (P@k) and nDCG@k tend to favor frequent labels, since correctly predicting head labels contributes equally to the score as predicting rare labels. As a result, these metrics may underestimate a model's ability to retrieve informative but infrequent labels.

Sweep	Value	nDCG@3	nDCG@5	P@1
Learning Rate (LR)	2×10^{-4}	0.619	0.680	0.880
	5×10^{-4}	0.648	0.713	0.920
	1×10^{-3}	**0.668**	**0.734**	**0.965**
	2×10^{-3}	0.555	0.610	0.820
	5×10^{-3}	0.461	0.507	0.730
Batch Size (BS)	8	0.600	0.660	0.850
	16	0.643	0.707	0.918
	32	**0.668**	**0.734**	**0.965**
	64	0.555	0.610	0.820

Table 6: One-variable hyperparameter sensitivity analysis. Each sweep varies one parameter while others are fixed at optimal values (lr = 10^{-3}, batch size = 32, pw = N). Best nDCG@5 per sweep is **bolded**.

Backbone	Approach	F1	MRR	mAP
		Immunology		
PubMedBERT	Baseline	0.209 ± 0.019	0.870 ± 0.014	0.299 ± 0.017
PubMedBERT	Proposed	**0.257** ± 0.014	**0.892** ± 0.021	**0.339** ± 0.036
BioBERT	Baseline	0.218 ± 0.016	0.850 ± 0.025	0.297 ± 0.012
BioBERT	Proposed	**0.255** ± 0.030	**0.891** ± 0.029	**0.341** ± 0.039
ClinicalBERT	Baseline	0.185 ± 0.039	0.825 ± 0.029	0.276 ± 0.016
ClinicalBERT	Proposed	**0.236** ± 0.020	**0.863** ± 0.012	**0.307** ± 0.025
SapBERT	Baseline	0.409 ± 0.012	0.927 ± 0.038	0.464 ± 0.033
SapBERT	Proposed	**0.435** ± 0.023	**0.971** ± 0.026	**0.507** ± 0.031
SciBERT	Baseline	0.240 ± 0.040	0.853 ± 0.037	0.317 ± 0.020
SciBERT	Proposed	**0.263** ± 0.013	**0.895** ± 0.012	**0.340** ± 0.017
		Embryology		
PubMedBERT	Baseline	0.315 ± 0.013	0.959 ± 0.025	0.408 ± 0.038
PubMedBERT	Proposed	**0.350** ± 0.038	**0.968** ± 0.031	**0.455** ± 0.014
BioBERT	Baseline	0.314 ± 0.016	0.941 ± 0.032	0.412 ± 0.011
BioBERT	Proposed	**0.348** ± 0.024	**0.972** ± 0.039	**0.459** ± 0.012
ClinicalBERT	Baseline	0.272 ± 0.015	0.906 ± 0.012	0.370 ± 0.016
ClinicalBERT	Proposed	**0.317** ± 0.027	**0.941** ± 0.040	**0.404** ± 0.015
SapBERT	Baseline	0.412 ± 0.037	0.952 ± 0.040	0.478 ± 0.034
SapBERT	Proposed	**0.443** ± 0.031	**0.973** ± 0.013	**0.521** ± 0.036
SciBERT	Baseline	0.312 ± 0.013	0.930 ± 0.012	0.401 ± 0.019
SciBERT	Proposed	**0.351** ± 0.028	**0.964** ± 0.020	**0.443** ± 0.016
		Neurology		
PubMedBERT	Baseline	0.373 ± 0.033	0.969 ± 0.018	0.437 ± 0.025
PubMedBERT	Proposed	**0.412** ± 0.025	**0.970** ± 0.028	**0.463** ± 0.022
BioBERT	Baseline	0.379 ± 0.040	0.968 ± 0.030	0.429 ± 0.012
BioBERT	Proposed	**0.406** ± 0.014	**0.971** ± 0.039	**0.466** ± 0.031
ClinicalBERT	Baseline	0.367 ± 0.034	0.969 ± 0.012	0.436 ± 0.025
ClinicalBERT	Proposed	**0.395** ± 0.035	0.968 ± 0.012	**0.452** ± 0.040
SapBERT	Baseline	0.413 ± 0.038	0.969 ± 0.023	0.475 ± 0.016
SapBERT	Proposed	**0.435** ± 0.021	**0.971** ± 0.013	**0.507** ± 0.016
SciBERT	Baseline	0.385 ± 0.031	0.969 ± 0.033	0.434 ± 0.023
SciBERT	Proposed	**0.407** ± 0.030	**0.972** ± 0.032	**0.471** ± 0.027

Table 7: Supplementary evaluation results for F1, MRR, and mAP across datasets. Values report mean with standard deviation (±).

To better evaluate performance under long-tailed label distributions, prior work (Jain et al., 2016; Gupta et al., 2021; Mittal et al., 2021; Saini et al., 2021; Wei et al., 2021; You et al., 2019b) proposes propensity-scored metrics that explicitly assign higher rewards to rare labels. Each label l is associated with an inverse propensity score defined as

$$\frac{1}{p_l} = 1 + C(N_l + B)^{-A},$$

where N_l denotes the number of training documents containing label l, and A, B, and C are dataset-dependent constants. Following established practice in extreme multi-label classification (Jain et al., 2016; Gupta et al., 2021; Mittal et al., 2021; Saini et al., 2021; Wei et al., 2021; You et al., 2019b) , we set $A = 0.55$, $B = 1.5$, and $C = (\log |D| - 1)(B + 1)^A$.

Using these weights, Propensity-Scored Precision (PSP@k) evaluates the top-k predictions by rewarding correct predictions of infrequent labels more strongly:

$$\text{PSP@k} = \frac{1}{k} \sum_{i=1}^{k} \frac{y_{d,\text{rank}(i)}}{p_{d,\text{rank}(i)}}.$$

Similarly, Propensity-Scored Discounted Cumulative Gain (PSDCG@k) extends DCG by incorporating the same propensity weights:

$$\text{PSDCG@k} = \sum_{i=1}^{k} \frac{y_{d,\text{rank}(i)}}{p_{d,\text{rank}(i)} \log(i+1)}.$$

The normalized metric PSN@k is then computed by dividing PSDCG@k by the maximum achievable value:

$$\text{PSN@k} = \frac{\text{PSDCG@k}}{\sum_{i=1}^{\min(k,|y_d|)} \frac{1}{\log(i+1)}}.$$

These metrics therefore provide a more faithful assessment of model behavior in long-tailed biomedical label spaces by emphasizing the correct retrieval of rare and semantically informative labels.

A.6 Hyperparameter Sensitivity Analysis

We analyze the sensitivity of the proposed model to two key training hyperparameters: the learning rate and the batch size. Table 6 reports the performance variation using three evaluation metrics: P@1, nDCG@3, and nDCG@5.

Learning Rate Sensitivity. Table 6 shows the effect of varying the learning rate from 2×10^{-4} to 5×10^{-3}. Performance improves as the learning rate increases from 2×10^{-4} to 1×10^{-3}, where the model achieves the best performance across all evaluation metrics. When the learning rate exceeds this value, performance drops noticeably, indicating that overly large learning rates destabilize optimization and negatively affect ranking quality. These results suggest that a learning rate of 1×10^{-3} provides the best balance between convergence speed and training stability.

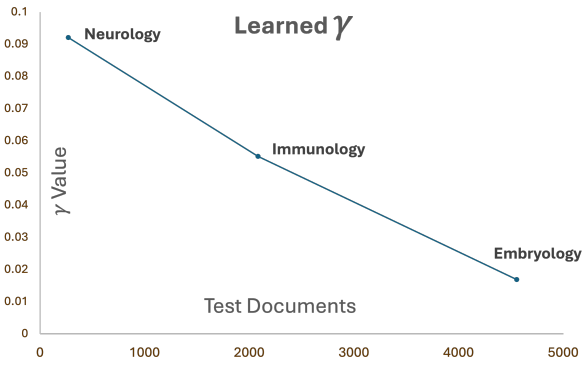


Figure 5: Correction Parameter Learning Curve over the domains

Batch Size Sensitivity. Table6 also reports the effect of different batch sizes. Increasing the batch size from 8 to 32 consistently improves P@1, nDCG@3, and nDCG@5, indicating that moderately larger batches help stabilize gradient updates and improve representation learning. However, further increasing the batch size to 64 results in a decline in performance, suggesting that excessively large batches may reduce gradient diversity and negatively affect generalization. Overall, a batch size of 32 achieves the best performance across the reported metrics.

These observations indicate that the proposed method remains relatively stable across a range of hyperparameter settings while achieving optimal performance with a learning rate of 1×10^{-3} and a batch size of 32.

Correction Parameter Sensitivity. The correction scale γ is parameterized as $\text{softplus}(\log \hat{\gamma})$ and learned jointly during training. Empirically, γ exhibits an inverse relationship with spurious direction density—the proportion of labels with valid directions as shown in Figure 5. In Neurology, where only 2.4% of labels are active, γ converges to 0.0921, whereas in Embryology, with 39.7% active labels, it reduces to 0.0169. This reflects an adaptive self-regularization effect: as more directions contribute per instance, γ shrinks to prevent over-correction. Learning γ thus enables automatic adjustment to domain-specific spuriousness without manual tuning.

We additionally evaluate a fixed-α global translation baseline, applying Equation 4 with a fixed scalar α in place of the learned γ and without the instance-specific gate $\tilde{s}_{i\ell}$, across $\alpha \in \{0.01, 0.02, 0.05, 0.10\}$ (Table 8).

	Neurology		Immunology		Embryology	
α	nDCG@3	nDCG@5	nDCG@3	nDCG@5	nDCG@3	nDCG@5
0.01	0.63	0.56	0.47	0.42	*0.72*	*0.61*
0.02	0.66	0.60	0.50	0.45	0.67	0.58
0.05	0.70	0.62	*0.57*	*0.47*	0.62	0.55
0.10	*0.73*	*0.65*	0.52	0.46	0.60	0.52
SPUR-XMC	**0.81**	**0.72**	**0.67**	**0.59**	**0.82**	**0.73**

Table 8: Fixed-α global translation across domains. Bold denotes SPUR-XMC (best overall); italics denote the best fixed α per domain. SPUR-XMC outperforms every fixed-α setting, and the best-performing α shifts lower as active spurious directions increase, mirroring the learned γ's self-regularization (Appendix A.6).

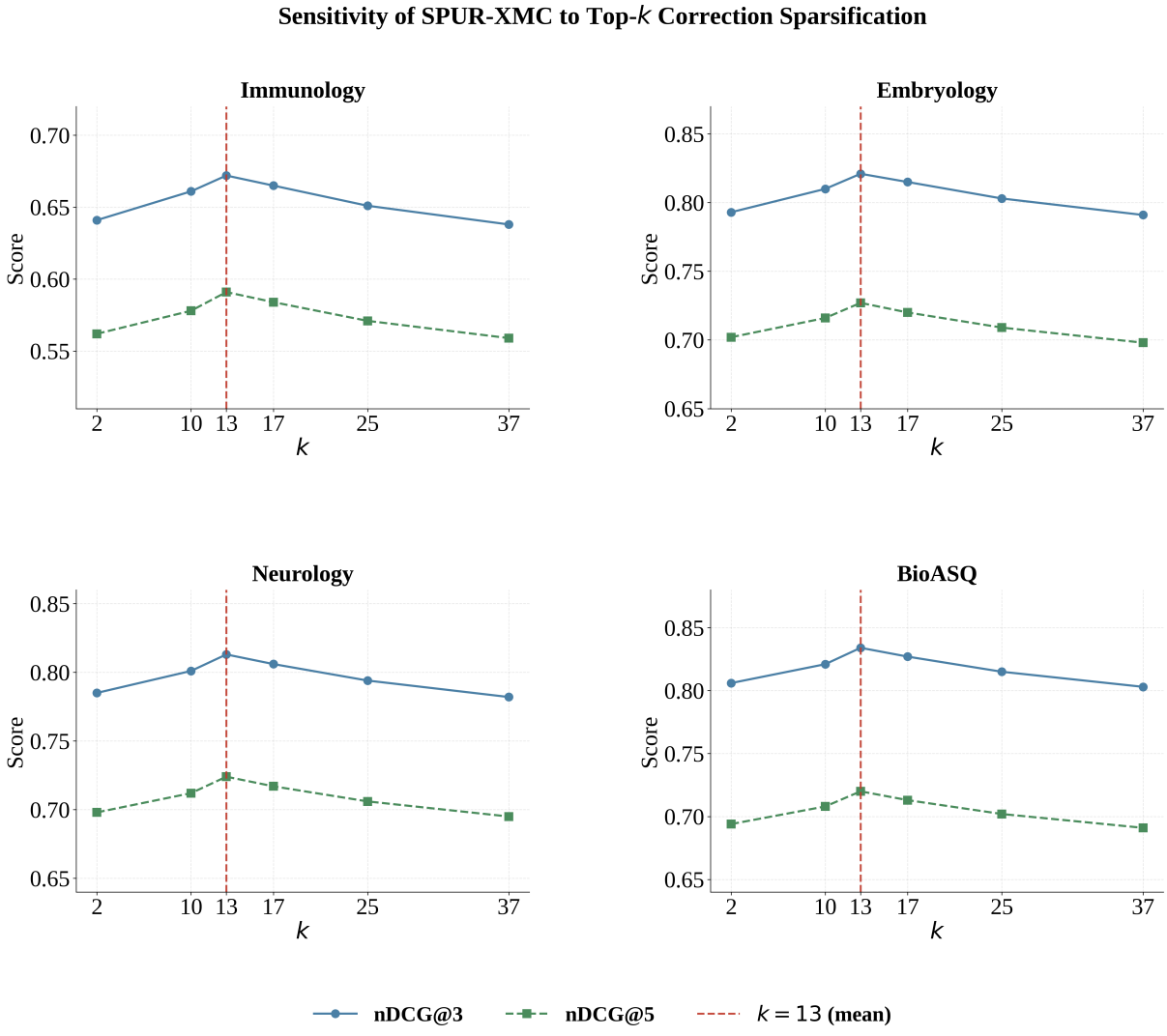


Figure 6: Sensitivity of SPUR-XMC to the top-k correction sparsification threshold across four biomedical domains. The red dashed line marks k=13, corresponding to the mean number of labels per document.

A.7 Additional Performance Evaluation

A.7.1 Supplementary Evaluation Results

Table 7 reports supplementary evaluation results in terms of F1, Mean Reciprocal Rank (MRR), and Mean Average Precision (mAP) across three biomedical domains—Immunology, Embryology, and Neurology—using five different biomedical pretrained encoders. Each backbone is evaluated under the baseline configuration and the proposed method, with values reported as mean ± standard deviation. Overall, the proposed method consistently improves performance across all backbones and datasets. In the Immunology domain, the pro-

Base Model	Method	P@1	P@3	P@5	nDCG@1	nDCG@3	nDCG@5
Immunology							
ClinicalBERT	Baseline	0.776 ± 0.031	0.559 ± 0.024	0.431 ± 0.011	0.760 ± 0.019	0.599 ± 0.017	0.502 ± 0.037
	Proposed	**0.803**† ± 0.031	**0.599**† ± 0.032	**0.487**† ± 0.022	**0.803**† ± 0.017	**0.643**† ± 0.032	**0.556**† ± 0.035
SapBERT	Baseline	0.931 ± 0.021	0.737 ± 0.019	0.631 ± 0.038	0.921 ± 0.027	0.770 ± 0.022	0.691 ± 0.022
	Proposed	**0.959**† ± 0.019	**0.775**† ± 0.013	**0.679**† ± 0.027	**0.959**† ± 0.035	**0.819**† ± 0.032	**0.750**† ± 0.037
SciBERT	Baseline	0.819 ± 0.014	0.595 ± 0.021	0.481 ± 0.030	0.799 ± 0.012	0.628 ± 0.014	0.533 ± 0.033
	Proposed	**0.842**† ± 0.013	**0.625**† ± 0.034	**0.521**† ± 0.022	**0.842**† ± 0.031	**0.673**† ± 0.028	**0.591**† ± 0.019
MedicalBERT	Baseline	0.870 ± 0.021	0.641 ± 0.018	0.534 ± 0.019	0.842 ± 0.017	0.683 ± 0.022	0.613 ± 0.024
	Proposed	**0.903**† ± 0.019	**0.689**† ± 0.016	**0.592**† ± 0.018	**0.903**† ± 0.021	**0.741** ± 0.019	**0.664** ± 0.020
Embryology							
ClinicalBERT	Baseline	0.881 ± 0.018	0.690 ± 0.031	0.555 ± 0.033	0.881 ± 0.019	0.714 ± 0.026	0.624 ± 0.039
	Proposed	**0.909**† ± 0.028	**0.724**† ± 0.025	**0.605**† ± 0.023	**0.909**† ± 0.033	**0.765**† ± 0.022	**0.673**† ± 0.035
SapBERT	Baseline	0.931 ± 0.022	0.784 ± 0.032	0.683 ± 0.029	0.931 ± 0.024	0.802 ± 0.037	0.712 ± 0.014
	Proposed	**0.954** ± 0.036	**0.822**† ± 0.023	**0.715**† ± 0.019	**0.954** ± 0.035	**0.853**† ± 0.039	**0.772**† ± 0.040
SciBERT	Baseline	0.921 ± 0.039	0.737 ± 0.017	0.610 ± 0.021	0.921 ± 0.018	0.759 ± 0.013	0.654 ± 0.028
	Proposed	**0.943** ± 0.029	**0.763**† ± 0.012	**0.645**† ± 0.032	**0.943** ± 0.029	**0.804**† ± 0.013	**0.712**† ± 0.038
MedicalBERT	Baseline	0.885 ± 0.020	0.797 ± 0.017	0.690 ± 0.018	0.885 ± 0.016	0.848 ± 0.023	0.740 ± 0.025
	Proposed	**0.925** ± 0.018	**0.840** ± 0.019	**0.725** ± 0.020	**0.925** ± 0.014	**0.881** ± 0.021	**0.795** ± 0.022
Neurology							
ClinicalBERT	Baseline	0.950 ± 0.021	0.762 ± 0.017	0.603 ± 0.026	0.950 ± 0.037	0.790 ± 0.014	0.697 ± 0.013
	Proposed	**0.959** ± 0.017	**0.766** ± 0.021	**0.627**† ± 0.036	**0.959** ± 0.010	**0.811**† ± 0.029	**0.710**† ± 0.023
SapBERT	Baseline	0.950 ± 0.016	0.766 ± 0.029	0.643 ± 0.035	0.950 ± 0.027	0.810 ± 0.016	0.727 ± 0.032
	Proposed	**0.959** ± 0.024	**0.775** ± 0.021	**0.679**† ± 0.022	**0.959** ± 0.017	**0.819** ± 0.014	**0.750**† ± 0.010
SciBERT	Baseline	0.948 ± 0.026	0.756 ± 0.035	0.613 ± 0.018	0.938 ± 0.028	0.796 ± 0.038	0.707 ± 0.027
	Proposed	**0.959** ± 0.030	**0.767** ± 0.037	**0.645**† ± 0.011	**0.959** ± 0.020	**0.813**† ± 0.034	**0.725**† ± 0.018
MedicalBERT	Baseline	0.948 ± 0.021	0.812 ± 0.018	0.693 ± 0.019	0.988 ± 0.017	0.868 ± 0.022	0.780 ± 0.024
	Proposed	**0.954** ± 0.019	**0.830**† ± 0.016	**0.711**† ± 0.018	**0.994** ± 0.015	**0.880**† ± 0.020	**0.795**† ± 0.021

Table 9: Performance comparison across three datasets for Remaining Encoders. Results are reported as mean ± standard deviation over 10 independent runs; bolded values indicate the best performance within each dataset, and †denotes statistically significant improvement over the baseline ($p < 0.05$).

posed approach yields notable gains over the baseline for every encoder. For instance, PubMedBERT improves from 0.209 to 0.257 in F1, from 0.870 to 0.892 in MRR, and from 0.299 to 0.339 in mAP. Similar trends are observed for BioBERT and ClinicalBERT, where the proposed method improves both ranking-based metrics (MRR, mAP) and classification performance (F1). SapBERT achieves the highest absolute scores in this domain indicating that the method remains effective even when applied to stronger biomedical encoders. Consistent improvements are also observed in the Embryology dataset. The proposed model improves F1 scores for all five backbones, with gains ranging from approximately 0.03 to 0.04. For example, SciBERT improves from 0.312 to 0.351 in F1 and from 0.401 to 0.443 in mAP. The largest absolute scores are again achieved with SapBERT, where the proposed method increases mAP from 0.478 to 0.521 and MRR from 0.952 to 0.973. These results suggest that the proposed approach enhances both retrieval quality and ranking accuracy across diverse biomedical contexts. A similar pattern is observed in the Neurology dataset. Across all encoders, the proposed method improves F1, MRR, and mAP relative to the baseline. For example, PubMedBERT shows an increase in F1 from 0.373 to 0.412 and in mAP from 0.437 to 0.463. BioBERT and SciBERT also exhibit consistent improvements across all metrics. SapBERT again achieves the strongest overall results, reaching an F1 of 0.435 and an mAP of 0.507 under the proposed configuration. Taken together, these results indicate that the proposed method provides consistent performance gains across multiple biomedical encoders and datasets. The improvements are observed not only in F1 but also in ranking-based metrics such as MRR and mAP, suggesting that the method enhances both label prediction accuracy and the quality of label ranking. The consistency of these gains across architectures further demonstrates that the proposed approach is encoder-agnostic and can ef-

Base Model	Method	PSP@1	PSP@3	PSP@5	PSN@3	PSN@5	PSP@1/P@1
Immunology							
PubMedBERT	Baseline	0.575 ± 0.020	0.603 ± 0.018	0.620 ± 0.019	0.590 ± 0.017	0.604 ± 0.018	0.71
	Proposed	**0.608** ± 0.017	**0.639** ± 0.016	**0.658** ± 0.017	**0.626** ± 0.015	**0.642** ± 0.016	0.72
BioBERT	Baseline	0.566 ± 0.021	0.594 ± 0.019	0.612 ± 0.020	0.580 ± 0.018	0.596 ± 0.019	0.70
	Proposed	**0.598** ± 0.018	**0.631** ± 0.017	**0.649** ± 0.018	**0.618** ± 0.016	**0.635** ± 0.017	0.72
ClinicalBERT	Baseline	0.551 ± 0.022	0.579 ± 0.020	0.596 ± 0.021	0.568 ± 0.018	0.583 ± 0.019	0.71
	Proposed	**0.582** ± 0.019	**0.614** ± 0.018	**0.632** ± 0.019	**0.603** ± 0.017	**0.619** ± 0.018	0.72
SapBERT	Baseline	0.624 ± 0.019	0.648 ± 0.017	0.664 ± 0.018	0.638 ± 0.016	0.651 ± 0.017	0.67
	Proposed	**0.651** ± 0.016	**0.683** ± 0.015	**0.702** ± 0.016	**0.670** ± 0.014	**0.686** ± 0.015	0.68
SciBERT	Baseline	0.582 ± 0.020	0.607 ± 0.018	0.623 ± 0.019	0.594 ± 0.017	0.607 ± 0.018	0.71
	Proposed	**0.611** ± 0.017	**0.642** ± 0.016	**0.660** ± 0.017	**0.629** ± 0.015	**0.644** ± 0.016	0.73
Embryology							
PubMedBERT	Baseline	0.623 ± 0.020	0.648 ± 0.018	0.662 ± 0.019	0.636 ± 0.017	0.650 ± 0.018	0.67
	Proposed	**0.652** ± 0.017	**0.686** ± 0.016	**0.705** ± 0.017	**0.673** ± 0.015	**0.690** ± 0.016	0.69
BioBERT	Baseline	0.620 ± 0.021	0.645 ± 0.019	0.659 ± 0.020	0.634 ± 0.018	0.648 ± 0.019	0.69
	Proposed	**0.662** ± 0.018	**0.695** ± 0.017	**0.714** ± 0.018	**0.682** ± 0.016	**0.699** ± 0.017	0.69
ClinicalBERT	Baseline	0.617 ± 0.022	0.640 ± 0.020	0.654 ± 0.021	0.628 ± 0.019	0.642 ± 0.020	0.70
	Proposed	**0.645** ± 0.019	**0.678** ± 0.018	**0.697** ± 0.019	**0.664** ± 0.017	**0.681** ± 0.018	0.71
SapBERT	Baseline	0.606 ± 0.020	0.629 ± 0.018	0.643 ± 0.019	0.618 ± 0.017	0.631 ± 0.018	0.65
	Proposed	**0.631** ± 0.017	**0.664** ± 0.016	**0.682** ± 0.017	**0.652** ± 0.015	**0.668** ± 0.016	0.66
SciBERT	Baseline	0.654 ± 0.021	0.677 ± 0.020	0.691 ± 0.021	0.666 ± 0.018	0.678 ± 0.019	0.71
	Proposed	**0.683** ± 0.018	**0.717** ± 0.017	**0.735** ± 0.018	**0.705** ± 0.016	**0.721** ± 0.017	0.72
Neurology							
PubMedBERT	Baseline	0.626 ± 0.019	0.649 ± 0.018	0.662 ± 0.019	0.639 ± 0.017	0.651 ± 0.018	0.66
	Proposed	**0.635** ± 0.017	**0.659** ± 0.016	**0.673** ± 0.017	**0.649** ± 0.015	**0.662** ± 0.016	0.66
BioBERT	Baseline	0.644 ± 0.020	0.668 ± 0.019	0.681 ± 0.020	0.657 ± 0.018	0.670 ± 0.019	0.68
	Proposed	**0.652** ± 0.018	**0.676** ± 0.017	**0.689** ± 0.018	**0.665** ± 0.016	**0.678** ± 0.017	0.68
ClinicalBERT	Baseline	0.637 ± 0.020	0.661 ± 0.019	0.674 ± 0.020	0.651 ± 0.018	0.664 ± 0.019	0.67
	Proposed	**0.645** ± 0.018	**0.670** ± 0.017	**0.683** ± 0.018	**0.660** ± 0.016	**0.673** ± 0.017	0.67
SapBERT	Baseline	0.618 ± 0.019	0.642 ± 0.018	0.656 ± 0.019	0.631 ± 0.017	0.644 ± 0.018	0.65
	Proposed	**0.624** ± 0.017	**0.648** ± 0.016	**0.662** ± 0.017	**0.637** ± 0.015	**0.651** ± 0.016	0.65
SciBERT	Baseline	0.654 ± 0.021	0.677 ± 0.020	0.690 ± 0.021	0.665 ± 0.018	0.678 ± 0.019	0.69
	Proposed	**0.661** ± 0.018	**0.685** ± 0.017	**0.698** ± 0.018	**0.673** ± 0.016	**0.686** ± 0.017	0.69

Table 10: Propensity-scored retrieval performance across the datasets. Results are reported as mean ± standard deviation over 10 independent runs; Bolded values indicate the best performance within each dataset.

fectively complement a wide range of biomedical language models.

A.7.2 Results on Remaining Encoders

The remaining encoder-specific results and the corresponding SPUR-XMC in Table 9 follow the same experimental protocol used in the main results table. Consistent improvements are observed across nearly all pretrained language models and datasets, indicating that the proposed mitigation framework generalizes across diverse biomedical and clinical encoders. Since these models are pretrained on different data sources, including biomedical literature, clinical narratives, and mixed-domain medical corpora, the results suggest that SPUR-XMC functions as an effective and architecture-agnostic enhancement that can be integrated with a broad range of BERT-based pretrained language models.

A.7.3 Propensity-Scored Retrieval Results

Table 10 reports propensity-scored retrieval performance across Immunology, Embryology, and Neurology. SPUR-XMC consistently improves PSP@k and PSN@k across backbone encoders and domains, indicating enhanced retrieval performance for infrequent and tail labels. The improvements are particularly pronounced in Embryology, which exhibits a stronger long-tail distribution, while Neurology shows comparatively smaller gains due to a less severe imbalance and stronger baseline performance. We further analyze the ratio PSP@1/P@1, which reflects the average reward

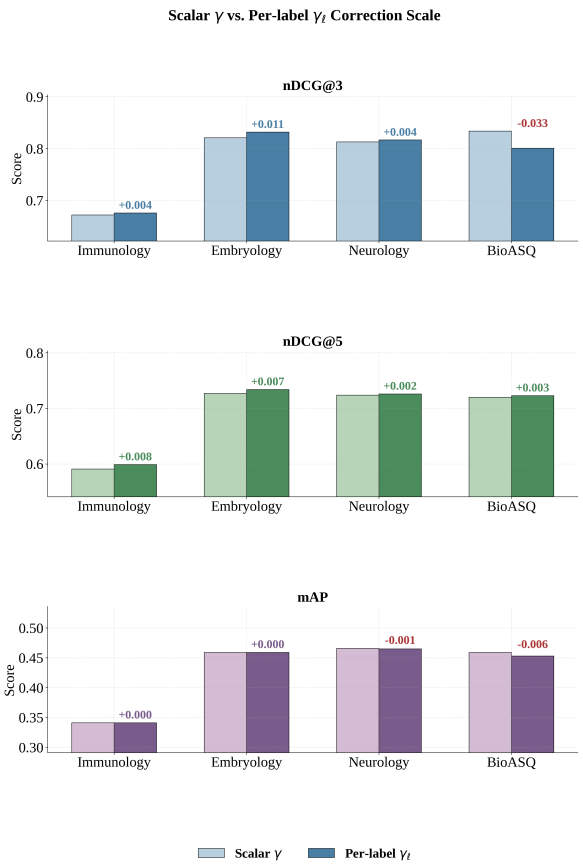


Figure 7: Comparison of scalar γ and per-label γ_ℓ correction scaling across four biomedical domains. Positive and negative deltas indicate performance gains and degradations of per-label γ_ℓ relative to scalar γ, respectively.

assigned to correctly predicted labels, with larger values indicating greater emphasis on less frequent concepts. Across encoders, the proposed method maintains ratios comparable to the baselines while consistently improving PSP@k and PSN@k, suggesting that SPUR-XMC improves overall ranking quality without sacrificing the retrieval of rare biomedical concepts.

A.8 Computational Cost Analysis

We report the practical cost of SPUR-XMC as an add-on module: (1) preprocessing time for CLS embedding collection and label-wise spurious direction estimation, (2) memory footprint of the direction matrix $V \in \mathbb{R}^{|\mathcal{L}^*| \times d}$, (3) trainable parameters relative to the backbone, and (4) per-batch inference overhead relative to the backbone encoder alone. Direction estimation is a one-time offline pre-computation that does not affect inference latency. Results across all four datasets are reported in Table 11.

A.9 Ablation Studies and Design Validation

A.9.1 Top-k sparsification sensitivity

Figure 6 examines the effect of varying the top-k sparsification threshold on nDCG@3 and nDCG@5 across four biomedical domains. The top-k mechanism retains only the k largest instance-specific correction weights $\tilde{s}_{i\ell}$ per document while suppressing the remaining corrections, preventing accumulated influence from numerous correlated label directions. The average number of labels per document ranges from 12.8 to 13.6 across datasets, and therefore $k = 13$ is selected as a consistent threshold across all domains. This setting achieves the strongest overall performance, whereas smaller values lead to under-correction and larger values introduce excessive corrections beyond the document's semantic scope. These findings suggest that the mean-based heuristic provides a principled and robust choice while maintaining stable performance across a broad range of k values.

A.9.2 Scalar versus Per-Label Correction Scaling

Figure 7 compares scalar γ and per-label γ_ℓ across four biomedical domains using nDCG@3, nDCG@5, and mAP. Per-label γ_ℓ provides only marginal improvements on Immunology, Embryology, and Neurology, with gains remaining below 0.011 across all metrics. On BioASQ, however, per-label γ_ℓ underperforms scalar γ on both nDCG@3 and mAP, suggesting that label-specific scaling introduces instability in large-scale settings where many labels have limited training observations. These findings indicate that the instance-specific gate $\tilde{s}_{i\ell}$ already captures label-level information through document–label activations, making an additional per-label scaling parameter largely redundant. We therefore adopt scalar γ as the default parameterization, providing improved stability and parameter efficiency without meaningful performance degradation.

A.9.3 Analysis of Direction Alignment

To empirically examine the conditions underlying Proposition 3.2, we compute $w_\ell^\top v_\ell$ for all seen labels using BioBERT encoders across three biomedical domains. Across Immunology, Embryology, and Neurology, the proportion of labels satisfying $w_\ell^\top v_\ell > 0$ ranges from 3.3% to 15.1%, confirming that the first failure mode described in Proposition 3.2, where uniform subtraction suppresses genuine semantic signal, occurs in practice. The

Domain	Preproc.	V (MB)	Params	% Backbone	Overhead
Neurology	3.0 min	5.4	2.0M	1.9%	0.49%
Immunology	1.5 min	21.1	6.1M	5.7%	0.24%
Embryology	4.2 min	33.6	9.4M	8.7%	0.24%
BioASQ	117 min	74.8	20.2M	18.7%	0.40%

Table 11: Computational cost of SPUR-XMC as an add-on module. Preproc. is one-time offline preprocessing time (CLS embedding collection plus direction estimation). % Backbone is trainable parameters as a percentage of PubMedBERT's $\sim$108M parameters. Overhead is full SPUR-XMC latency relative to the backbone-only forward pass (batch size 32).

remaining labels exhibit $w_\ell^\top v_\ell < 0$, indicating that the classifier tends to oppose the estimated spurious direction, yet uniform subtraction still distorts these associations by increasing classification scores regardless of semantic grounding. The co-existence of both alignment patterns across domains indicates substantial heterogeneity within the label space, suggesting that no fixed scalar α_ℓ can provide consistent mitigation across labels. This observation directly motivates the document- and label-specific correction weight $\tilde{s}_{i\ell}$ introduced in Section 3.2.3.